Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-09 | September - 2025 | Page 38-49

Preprocessing-Leakage Control, Repeated Internal Validation and Calibration Assessment in Structured Heart-Disease Classification

[1]Rakesh Kumar Khillan, [2]Dr. Abhinav Shukla
[1]Research Scholar, [2]Associate Professor
[1,2]Department of IT & CS, Dr. C. V. Raman University, Bilaspur, India
[1]rakeshkhillan@gmail.com, [2]amitshukla9611@gmail.com

Abstract

Clinical machine-learning studies can yield optimistic and clinically incomplete results if preprocessing is done outside the validation boundary, if the performance of the model is summarized on one split or if there is no probability calibration. In this work we investigated a framework for binary evaluation of recorded heart-disease status, which involves the preprocessing of leakage control, as well as explicitly stating the limitations of non-nested evolutionary feature selection. Methods: Public data for 918 records and 11 predictors were analyzed. A preliminary 20-fold stratified benchmark was conducted to evaluate Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes and Artificial Neural Network models. The common model family was then random forest, followed by the selected configurations by both the genetic algorithm (GA) and the particle swarm optimization (PSO). The learned operations were worked out on training folds and applied to validation folds. A 10-repetition 20-fold stratified cross-validation was performed with the fixed Random Forest configurations. The accuracy, precision, sensitivity, F1 score, ROC-AUC, exploratory paired tests and calibration of the compact GA-RF was reported. Conclusion: In the first comparison, SVM was the best non-RF one (accuracy 86.17%; ROC-AUC 0.9216). In repeated evaluation, full-feature Random Forest achieved accuracy of 87.11% ± 5.06% and ROC-AUC of 0.9285 ± 0.0396, compared with 83.67% ± 5.45% and 0.9075 ± 0.0439 for GA-RF and 80.74% ± 4.94% and 0.8843 ± 0.0418 for PSO-RF. The nominal fold-level paired tests highlighted the full-feature model, but were considered exploratory given that the repeated-cross-validation fold levels are correlated. The GA-RF Brier score was 0.1181; compared to an approximate prevalence-only Brier score of 0.2472, that yields a GA-RF internal Brier skill score of ~0.52 if the two scores are applied to the same prediction population. Findings: The fold-contained preprocessing and repeated resampling, probability assessment and evidence-aware interpretation make a significant difference in internal evaluation. But fixed masks, which are picked prior to repeated comparison are not complete nested feature selection. External validation and nested model development is still required prior to clinical-use claims.

Keywords: data leakage; repeated cross-validation; calibration; Brier score; random forest; heart-disease classification; clinical machine learning

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-09 | September - 2025 | Page 38-49

1. Introduction

With structured clinical data, machine learning algorithms are now used to analyze the data and if there are methodological errors, it is hard to believe in the results as they can seem strong. Imputation, encoding, scaling, feature selection or tuning may be conducted before the split into training and evaluation data has been made, which can lead to information leakage if properties of the evaluation data affect the modeling process [1]. The second issue is the lack of using multiple train-test splits or multiple cross-validation schedules, which may mask partition dependent variability. The third is discrimination or classification metrics without looking at whether the predicted probabilities are in line with outcome frequencies [2-5].

In moderately-sized public data sets these issues are particularly relevant. Using small validation folds yields different estimates of the metrics because there is no replication and using a large amount of data in repeated cross-validation causes data to be statistically dependent. Reporting hundreds of fold scores does not mean hundreds of independent experiments! When the nominal p value is small, but the effective uncertainty is larger, the resulting uncorrected paired tests will have very small p values. Transparent analysis requires therefore to differentiate between two types of repeated-validation evidence: descriptive (repeated validation) and formal independent replication.

Several reporting and risk of bias frameworks have a focus on explicit definitions of outcomes, preprocessing boundaries, model development procedures, calibration and applicability [2,3,16-18]. These principles apply to classical algorithms as well as to ensembles based on machine learning. A good model can miscalibrate and a good Brier score in one internal sample does not guarantee good calibration in another hospital, another time or another prevalence sample [4,5].

The Heart Failure Prediction Dataset used in this study is publicly available and the outcome column is HeartDisease [15]. It is thus a binary classification problem, not prospect prediction of heart failure. This distinction allows a cross sectional classifier to not be confused with a future risk or a time to event model.

This paper aims to evaluate an evidence-hierarchical comparison of the benchmark classifiers and three configurations of the Random Forest. The benchmark stage is a set of algorithms that offers algorithmic context. The principle, internal comparison is provided by the repeated Random Forest analysis. For the compact model selected by the calibration, the assessment of the model is reported. Importantly, the study does not claim to be fully leakage-free feature selection, as the leakage masking was performed using GA and PSO masks and these masks were not rediscovered during each outer training split. Instead, the goal is to show clear, non-deployment-limiting evaluation and limitations, rather than that they would be ready to deploy.

2. Materials and methods

2.1 Dataset and target definition

The data set consisted of 918 records, 11 structured predictors and a binary target HeartDisease [15]. Predictor set was Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak and ST_Slope. There were 410

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-09 | September - 2025 | Page 38-49

records in class 0 and 508 records in class 1. Class 1 was used as the positive class. The class distribution was moderately balanced, thus no synthetic oversampling or undersampling was done.

2.2 Preprocessing-leakage control

The target was excluded from the Predictor matrix during Preprocessing. Numerical standardization was performed on models sensitive to numerical values and a consistent training-derived representation was used to encode categorical variables. Any imputation, encoding or scaling operation which estimated parameters from the data was fitted on the training fold and applied without refitting to the validation fold. This is training/validation separation, which is the main protection from preprocessing leakage [1].

There were no assumptions that tree-based models need to be standardized. Using standardized numerical inputs, Logistic Regression, Support Vector Machine and Artificial Neural Network models were used. The aim of the methodology was not to implement the same changes to all algorithms, but to process the information in the same way, which is appropriate to the model.

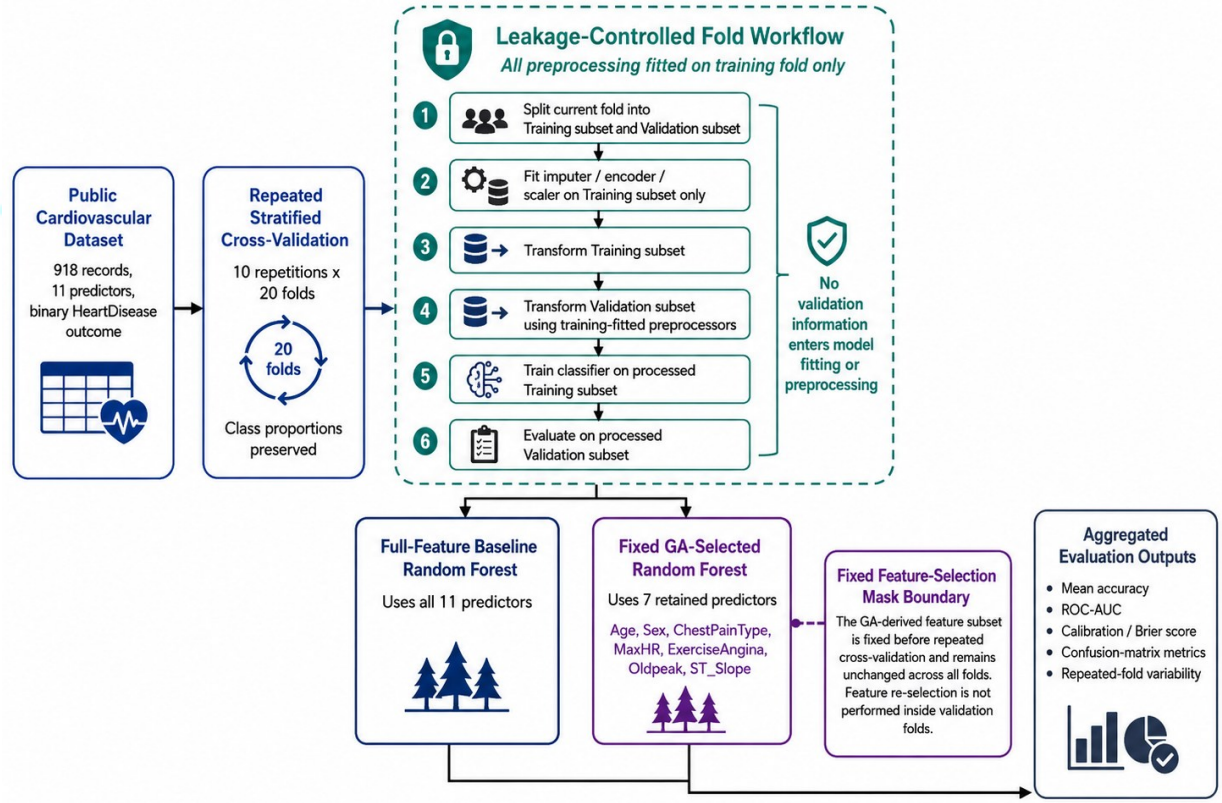

Figure 1. Preprocessing-leakage-controlled validation workflow and the stated boundary of fixed feature-selection masks.

2.3 Benchmark classifiers

The first benchmarking stage comprised of Logistic Regression, Decision Tree, Support Vector Machine (SVC), Gaussian Naïve Bayes and a feed-forward Artificial Neural Network (ANN). These algorithms were linear, rule-based, margin-based, probabilistic and connectionist learning paradigms. One 20-fold stratified cross-validation schedule was used to evaluate them.

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-09 | September - 2025 | Page 38-49

The benchmark comparison is descriptive, as the models do not have an exhaustive hyperparameter search budget.

Table 1. Benchmark model configuration and preprocessing requirements.

Model	Key configuration	Preprocessing emphasis
Logistic Regression	L2 penalty; lbfgs; C = 1.0	Fold-contained encoding and scaling
Decision Tree	Gini; random state 42	Fold-contained encoding; no scaling requirement
Support Vector Machine	RBF kernel; C = 1.0; gamma = scale	Fold-contained encoding and scaling
Naïve Bayes	GaussianNB; var_smoothing = 1×10^{-9}	Consistent numerical representation
Artificial Neural Network	One hidden layer; ReLU; Adam	Fold-contained encoding and scaling
Random Forest	100 trees; Gini; max_features = sqrt	Fold-contained encoding; no scaling requirement

2.4 Random Forest configurations and feature-selection boundary

The full-feature Random Forest was used to include all 11 predictors and was the major predictive reference. GA-RF took the following 7 variables: Age, Sex, ChestPainType, MaxHR, ExerciseAngina, Oldpeak and ST_Slope. PSO-RF included Age, ChestPainType, RestingBP, Cholesterol, FastingBS, ExerciseAngina and Oldpeak.

The GA and PSO masks were produced using internal Random Forest wrapper searches and then were fixed prior to repeated comparison. The GA required 20 chromosomes and 10 generations, while PSO required 30 particles and 50 iterations. The performance of the selected model may have selection-induced optimism as feature discovery was not repeated separately within each outer fold. The design should be called "preprocessing-leakage controlled with non-nested feature selection", rather than "fully leakage-free development pipeline".

2.5 Validation hierarchy

Three levels of evidence were all kept. The benchmarks models were tested in the first run with a 20-fold stratified cross-validation analysis. Second, a preliminary 20-fold comparison of the three Random Forest configurations was kept to show partition sensitivity. Third, for the principal Random Forest comparison, the cross-validation was repeated 10 times with 20-fold stratified CV which resulted in 200 cross-validation fold-level results per model.

The repeated design minimized the reliance on one partition, but the folds were correlated due to training observations being repeated. There was, therefore, a focus on mean and standard deviation. For completeness, the results of the nominal paired t-test and Wilcoxon signed-rank test are reported, but should be viewed as exploratory. The inferential evidence provided by a corrected resampled test or a repetition-level analysis or a paired bootstrap of out-of-fold predictions or a fully nested external evaluation would be stronger.

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-09 | September - 2025 | Page 38-49

Table 2. Validation hierarchy and evidential role.

Analysis	Models	Purpose	Interpretive status
Initial 20-fold benchmark	LR, DT, SVM, NB, ANN	Algorithmic context	Secondary internal evidence
Preliminary 20-fold RF comparison	Full RF, GA-RF, PSO-RF	Illustrate one partition schedule	Preliminary; not decisive
Repeated 10 × 20-fold comparison	Full RF, GA-RF, PSO-RF	Principal internal mean and variability	Primary internal evidence
GA-RF calibration	GA-RF probabilities	Internal probability-reliability assessment	Not comparative or external

2.6 Performance and calibration measures

Accuracy, precision, sensitivity, F1-score and ROC-AUC were calculated. The accuracy is the measure of overall correctness of the classification; precision is related to the correctness of the positive classes; sensitivity is the measure of the correctness of the positive records; F1-score is a compromise between the correctness of the positive classes and the correctness of the positive records; ROC-AUC is a measure of the ranking discrimination over the different probability thresholds [8-10]. None of the measures was used as a comprehensive definition of clinical utility.

The reliability of GA-RF was evaluated with a reliability curve and the Brier score: $BS = \frac{1}{N}\sum_{i=0}^{N}(p_i - y_i)^2$, where p_i is the predicted probability and y_i is the observed binary outcome. A prevalence only reference with a positive proportion of the observed samples of size p has Brier score of $p(1-p)$. A skill score was created as a ratio, Brier skill $= (1 - \frac{BSmodel}{BSreference})$. The squared probability error for a positive BSS is smaller than the prevalence-only reference. When comparing to a reference score, the comparison is only valid if the model and reference scores are computed on the same prediction population [4,5].

3. Results

3.1 Initial benchmark performance

In the first 20-fold comparison, Support Vector Machine (SVM) performed best with an accuracy of 86.17%, a sensitivity of 90.17%, an F1-score of 87.79% and an ROC-AUC of 0.9216. The Naïve Bayes and Logistic Regression methods also achieved comparable discrimination. Decision Tree performed the least well in the benchmarks. These results clearly show that there are several effective families of classifiers available on this dataset and the choice of Random Forest for the evolutionary analysis is not only motivated by the fact that the initial benchmark was optimal, but is also motivated by methodological suitability.

Table 3. Initial 20-fold benchmark classifier performance.

Model	Accuracy	Precision	Sensitivity	F1-score	ROC-AUC
Logistic Regression	85.08%	86.10%	87.78%	86.59%	0.9097

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-09 | September - 2025 | Page 38-49

Decision Tree	77.34%	81.25%	77.34%	78.97%	0.7733
Support Vector Machine	86.17%	85.95%	90.17%	87.79%	0.9216
Naïve Bayes	84.43%	86.55%	86.01%	85.91%	0.9192
Artificial Neural Network	82.69%	84.97%	84.06%	84.14%	0.9003

3.2 Preliminary and repeated Random Forest results

The preliminary results are presented in 20-fold comparison, where the accuracies of the full-feature RF, GA-RF and PSO-RF are 83.11%, 83.76% and 81.37%, respectively. This would appear to indicate, when taken alone, a marginal GA advantage. The overall repeated analysis changed that; the full-feature random forest outperformed all others in terms of mean accuracy, precision, sensitivity, F1 score and mean ROC-AUC.

The difference between the first analysis and second shows the importance of not making model superiority judgements based on a single partition schedule. The initial values are kept as evidence of Partition sensitivity and the repeated values are the main conclusion of the interior.

Table 4. Repeated 10 × 20-fold performance of Random Forest configurations.

Configuration	Accuracy, %	Precision, %	Sensitivity, %	F1-score, %	ROC-AUC
Full-feature RF	87.11 ± 5.06	86.99 ± 5.57	90.58 ± 6.05	88.59 ± 4.52	0.9285 ± 0.0396
GA-selected RF	83.67 ± 5.45	84.96 ± 5.79	86.04 ± 6.71	85.33 ± 4.94	0.9075 ± 0.0439
PSO-selected RF	80.74 ± 4.94	83.01 ± 5.54	82.49 ± 6.98	82.52 ± 4.60	0.8843 ± 0.0418

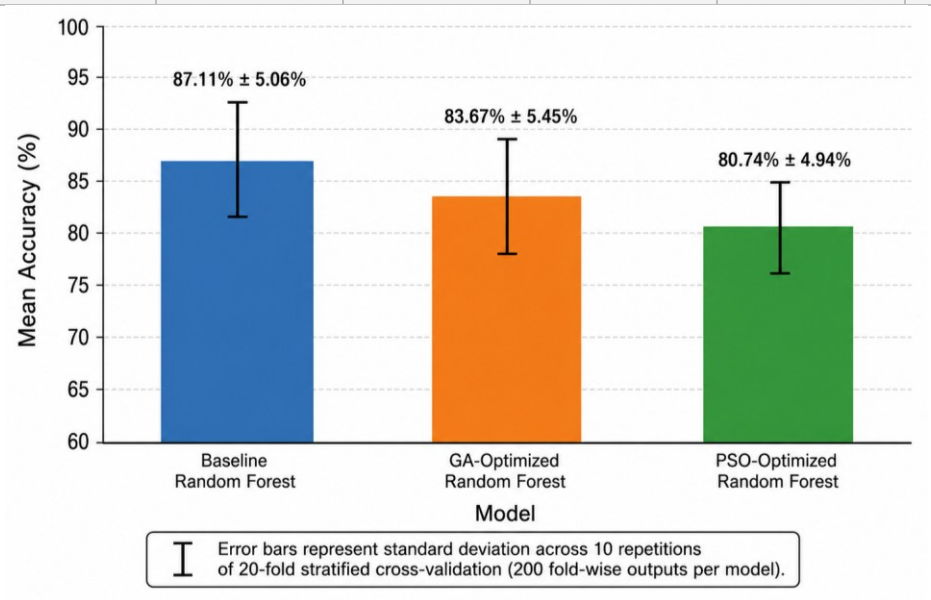

Figure 2. Mean accuracy and fold-level standard deviation under repeated 10 × 20-fold internal validation.

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-09 | September - 2025 | Page 38-49

3.3 Exploratory paired comparison

GA-RF minus full-feature RF had a mean difference of −3.44 percentage point with a reported nominal 95% interval of −4.00 to −2.88 for the 200 matched fold-level accuracy values. The Shapiro-Wilk test indicated non-normality ($W = 0.9710$, $p < 0.001$). The paired t-test gave a t value of −12.0419 and the Wilcoxon test gave a W value of 1517 with nominal p-values both less than 0.001. The 200 observations were from repeated, overlapping folds and were not considered as conclusive evidence of statistical superiority.

Table 5. Exploratory fold-level comparison of GA-RF and full-feature RF.

Measure	Value	Interpretation
Number of matched folds	200	Correlated repeated-CV outputs
Mean difference (GA − full)	−3.44 percentage points	Direction favours full-feature RF
Nominal 95% interval	−4.00 to −2.88	May be too narrow under fold dependence
Shapiro-Wilk	$W = 0.9710$; $p < 0.001$	Paired differences not normally distributed
Paired t-test	$t = -12.0419$; $p < 0.001$	Exploratory
Wilcoxon signed-rank	$W = 1517$; $p < 0.001$	Exploratory

3.4 Calibration context for the compact GA-RF

The GA-RF Brier score was 0.1181. The observed prevalence class-1 was 508/918 = 0.5534 with an approximate prevalence-only Brier score of 0.2472. The resulting Brier skill score for the two was around 0.52, if both scores were computed on the same prediction population. Therefore, GA-RF decreased by approximately 52% the squared probability error from that of giving identical prevalence probability for all records.

The calibration curve was generally parallel to the y=x line with some local deviations. This result suggests promising internal probability information, but not information from one institution to another or from one prevalence setting to another. No conclusion can be drawn that the GA-RF was best-calibrated, as calibration was not reported comparably for the full-feature and PSO configurations.

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-09 | September - 2025 | Page 38-49

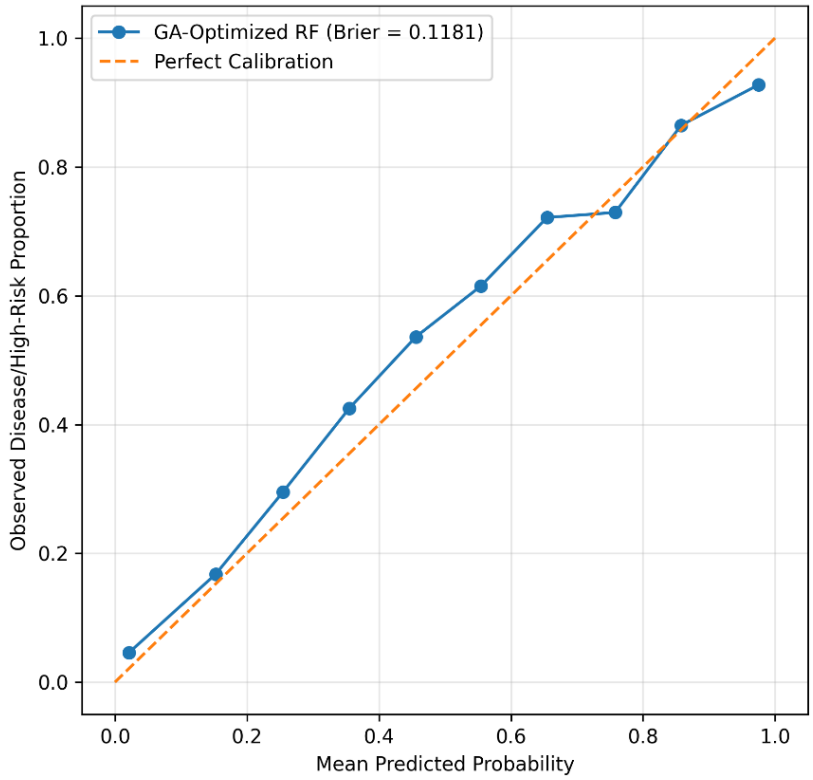

Figure 3. Internal calibration curve of the compact GA-selected Random Forest (reported Brier score = 0.1181).

Table 6. Contextual interpretation of the GA-RF Brier score.

Quantity	Value	Interpretation
GA-RF Brier score	0.1181	Internal squared probability error
Observed prevalence	0.5534	Class-1 proportion
Prevalence-only Brier reference	≈ 0.2472	p(1−p)
Approximate Brier skill score	≈ 0.52	Positive internal skill versus prevalence-only prediction

4. Discussion

4.1 Why preprocessing boundaries matter

The first methodological contribution is that of separating pre-processing of training data from validation data. Distributional information can also be transferred with even transformations that do not involve the target, provided they are fitted globally. The magnitude of the resulting optimism depends on the different datasets, but whether or not the numerical effect that is observed is large or not is not the correct design. Learners' encoders, imputers, scalers should be placed within the training process [1].

Leakage control is not a simple on/off characteristic. The current framework is able to handle pre-processing leakage; but fixed GA and PSO masks were given the dataset prior to re-comparison. This selection boundary is not a completely nested pipeline. It would therefore be

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-09 | September - 2025 | Page 38-49

incorrect to say the whole experiment was “strictly leakage-free”. The latter would be better described as a non-nested feature selection based on internal validation which is controlled by preprocessing.

4.2 Repeated validation and evidence hierarchy

The preliminary 20-fold analysis had a 0.65 percent-point lead for GA-RF while the repeated analysis had a 3.44 percent-point lead for the full-feature model. This turnaround is the most obvious empirical evidence for an evidence hierarchy. A model is not necessarily superior based on one favorable partition scheduling, especially if the folds are small.

The more cross-validation, the more uncertainty. Typical standard deviations of about 5 percentage points indicate that there is reasonable variation between folds. Furthermore, there is correlated score overlap in training sets. Reporting means ± SD is informative, but all 200 folds are not independent observations, causing one to overestimate the precision. Future confirmative work should be performed using corrected resampling (CRS), repetition-level comparisons, paired bootstrap of unified out-of-fold predictions or nested cross-validation.

4.3 Benchmark interpretation

The initial comparison evaluated it and the most competitive of the non-Random-Forest benchmark was Support Vector Machine. Random Forest would not be appropriate to state that it was empirical better than all other algorithms since the benchmark models were not run through the same repeated-analysis stream. However, the common evolutionary fitness model is chosen as the random forest model, which can be applied to structured data with nonlinear relationship and has the consistency in comparing full-versus-reduced [6,7]. This distinction is crucial from a methodological viewpoint and is different from numerical ranking without support.

4.4 Discrimination is not calibration

The GA-RF ROC-AUC value is 0.9075, suggesting that the internal ranking discrimination is high but the Brier score shows that its probability estimates are fairly accurate. While being informative, the positive Brier skill relative to a prevalence-only reference does not define calibration slope, calibration intercept, threshold-specific utility or transportability [4,5].

A thorough clinical prediction study should have similar out-of-fold probabilities for all candidate models, report uncertainty about Brier score and calibration parameters and consider recalibration for the models after external evaluation and assess net benefit over plausible thresholds [22,23]. The calibration analysis as it stands is thus supportive, but is not complete.

4.5 Implications for trustworthy medical AI

The study is an illustration of a general rule: methodological transparency is of more value than a slightly higher grade. The final conclusion altered with the addition of the evidence base from one cross validation schedule to repeated validation. Likewise, the paper also separates feature selection as a part of the preprocessing stage from nested feature selection and internal calibration from clinical utility. These differences align with the TRIPOD+AI, PROBAST+AI, MI-CLAIM and responsible clinical-AI recommendations [2,3,18-21].

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-09 | September - 2025 | Page 38-49

The GUI developed is just a research presentation layer and should not be considered as a final user's one. The external validation, calibration monitoring, subgroup and fairness evaluation, cyber security, auditability, human factors testing, workflow integration and regulatory review would be required if the system were to be deployed in the clinic [20,21,28].

5. Limitations

- A single moderate size public dataset with the label HeartDisease was used for the analysis.
- The repeated comparison was not conducted on the benchmark classifiers which were evaluated in one 20-fold schedule.
- Repeatedly evaluated, non-nested inside outer folds, GA and PSO masks were selected before.
- The repeated fold scores were correlated and the nominal paired-test p-values are exploratory.
- GA and PSO had different budget for candidate evaluation and a single search was reported.
- Only GA-RF model was evaluated in detail with regard to calibration and not compared with other models.
- The provenance of the separate 184-record GUI confusion matrix was not adequate for primary inference and was not a part of the central analysis.
- No external, prospective, subgroup, fairness, decision-curve and clinical-impact evaluation was conducted.

6. Conclusion

More than simply using a classifier on a public dataset is needed to have a trustworthy internal evaluation. For this study training-derived preprocessing was separated from the validation folds, one-schedule results were seconded by repeated resampling, dependence of the folds was revealed and discrimination was complemented by probability assessment. The three Random Forest configurations performed well; the full-feature Random Forest was the best one that was repeated. The compact GA-RF had a positive internal Brier skill and lower discrimination and PSO-RF even lower. The results justify minimum components of responsible analysis, namely preprocessing-leakage control, repeated validation and calibration assessment. They are not a substitute for the sequential, but important, process of complete nesting of features and independent external validation.

Declarations

Ethics statement: This was a secondary study of a de-identified, publicly available dataset and no new health information was collected or the study involved direct recruitment of participants.
Data access: Heart Failure Prediction Dataset is available open access from Kaggle [15].
Availability of the analysis code: The analysis code and configuration files should be deposited in a versioned public repository and are available from the authors upon reasonable request.

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-09 | September - 2025 | Page 38-49

References

[1] S. Kaufman, S. Rosset, C. Perlich, O. Stitelman, Leakage in data mining: Formulation, detection and avoidance, ACM Transactions on Knowledge Discovery from Data 6 (2012) 1-21. https://doi.org/10.1145/2382577.2382579.

[2] G.S. Collins, K.G.M. Moons, P. Dhiman, et al., TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ 385 (2024) e078378. https://doi.org/10.1136/bmj-2023-078378.

[3] K.G.M. Moons, J.A.A. Damen, T. Kaul, et al., PROBAST+AI: An updated quality, risk of bias and applicability assessment tool for prediction models using regression or artificial intelligence methods, BMJ 388 (2025) e082505. https://doi.org/10.1136/bmj-2024-082505.

[4] E.W. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation and Updating, 2nd ed., Springer, 2019. https://doi.org/10.1007/978-3-030-16399-0.

[5] B. Van Calster, D.J. McLernon, M. van Smeden, L. Wynants, E.W. Steyerberg, Calibration: The Achilles heel of predictive analytics, BMC Medicine 17 (2019) 230. https://doi.org/10.1186/s12916-019-1466-7.

[6] L. Breiman, Random forests, Machine Learning 45 (2001) 5-32. https://doi.org/10.1023/A:1010933404324.

[7] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273-297. https://doi.org/10.1007/BF00994018.

[8] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (2006) 861-874. https://doi.org/10.1016/j.patrec.2005.10.010.

[9] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management 45 (2009) 427-437. https://doi.org/10.1016/j.ipm.2009.03.002.

[10] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (2020) 6. https://doi.org/10.1186/s12864-019-6413-7.

[11] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995, pp. 1137-1145.

[12] D.G. Altman, P. Royston, What do we mean by validating a prognostic model?, Statistics in Medicine 19 (2000) 453-473.

[13] F.E. Harrell Jr., Regression Modeling Strategies, 2nd ed., Springer, 2015. https://doi.org/10.1007/978-3-319-19425-7.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825-2830.

[15] F. Soriano, Heart Failure Prediction Dataset, Kaggle, 2021. Retrieved June 5, 2026, from https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction.

[16] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G.M. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD), Annals of Internal Medicine 162 (2015) 55-63. https://doi.org/10.7326/M14-0697.

[17] R.F. Wolff, K.G.M. Moons, R.D. Riley, et al., PROBAST: A tool to assess the risk of bias and applicability of prediction model studies, Annals of Internal Medicine 170 (2019) 51-58. https://doi.org/10.7326/M18-1376.

[18] B. Norgeot, G. Quer, B.K. Beaulieu-Jones, et al., Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist, Nature Medicine 26 (2020) 1320-1324. https://doi.org/10.1038/s41591-020-1041-y.

[19] A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, New England Journal of Medicine 380 (2019) 1347-1358. https://doi.org/10.1056/NEJMra1814259.

[20] J. Wiens, S. Saria, M. Sendak, et al., Do no harm: A roadmap for responsible machine learning for health care, Nature Medicine 25 (2019) 1337-1340. https://doi.org/10.1038/s41591-019-0548-6.

[21] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Medicine 17 (2019) 195. https://doi.org/10.1186/s12916-019-1426-2.

[22] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach, Biometrics 44 (1988) 837-845.

[23] A.J. Vickers, E.B. Elkin, Decision curve analysis: A novel method for evaluating prediction models, Medical Decision Making 26 (2006) 565-574.

[24] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, et al., Explainable artificial intelligence: Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82-115.

[25] E. Tjoa, C. Guan, A survey on explainable artificial intelligence: Toward medical XAI, IEEE Transactions on Neural Networks and Learning Systems 32 (2021) 4793-4813.

[26] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157-1182.

[27] B. Xue, M. Zhang, W.N. Browne, X. Yao, A survey on evolutionary computation approaches to feature selection, IEEE Transactions on Evolutionary Computation 20 (2016) 606-626.

[28] E.H. Shortliffe, M.J. Sepúlveda, Clinical decision support in the era of artificial intelligence, JAMA 320 (2018) 2199-2200.

[29] A.L. Beam, I.S. Kohane, Big data and machine learning in health care, JAMA 319 (2018) 1317-1318.

[30] Z. Obermeyer, E.J. Emanuel, Predicting the future—Big data, machine learning and clinical medicine, New England Journal of Medicine 375 (2016) 1216-1219.

[31] Y. Cai, Y.-Q. Cai, L.-Y. Tang, et al., Artificial intelligence in cardiovascular risk prediction models: A systematic review, BMC Medicine 22 (2024) 56.